Genetical metabolomics of flavonoid biosynthesis in *Populus*: a case study

Kris Morreel[1], Geert Goeminne[1], Véronique Storme[1], Lieven Sterck[1], John Ralph[2], Wouter Coppieters[3], Peter Breyne[4], Marijke Steenackers[4], Michel Georges[3], Eric Messens[1] and Wout Boerjan[1,*]

[1]*Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, B-9052 Gent, Belgium,*
[2]*United States Dairy Forage Research Center, Agricultural Research Service, United States Department of Agriculture and Department of Forestry, University of Wisconsin, Madison, WI 53706, USA,*
[3]*Department of Genetics, Faculty of Veterinary Medicine, University of Liège, B-4000 Liège, Belgium, and*
[4]*The Research Institute for Nature and Forest (INBO), B-9500 Geraardsbergen, Belgium*

Summary

Genetical metabolomics [metabolite profiling combined with quantitative trait locus (QTL) analysis] has been proposed as a new tool to identify loci that control metabolite abundances. This concept was evaluated in a case study with the model tree *Populus*. Using HPLC, the peak abundances were analyzed of 15 closely related flavonoids present in apical tissues of two full-sib poplar families, *Populus deltoides* cv. S9-2 × *P. nigra* cv. Ghoy and *P. deltoides* cv. S9-2 × *P. trichocarpa* cv. V24, and correlation and QTL analysis were used to detect flux control points in flavonoid biosynthesis. Four robust metabolite quantitative trait loci (mQTL), associated with rate-limiting steps in flavonoid biosynthesis, were mapped. Each mQTL was involved in the flux control to one or two flavonoids. Based on the identities of the affected metabolites and the flavonoid pathway structure, a tentative function was assigned to three of these mQTL, and the corresponding candidate genes were mapped. The data indicate that the combination of metabolite profiling with QTL analysis is a valuable tool to identify control points in a complex metabolic pathway of closely related compounds.

Keywords: poplar, flavonoids, metabolite profiling, correlation analysis, mQTL, genetical metabolomics.

Introduction

Metabolite profiling has gained much attention as a powerful functional genomics tool to unravel gene function (Fiehn *et al.*, 2000; Goodacre *et al.*, 2004; Sumner *et al.*, 2003). For example, comparative metabolite profiling of wild-type and mutant plants has shown that mutations in single genes can affect the concentrations of a wide variety of metabolites (Rohde *et al.*, 2004). In addition, metabolite profiling of plants grown under various environmental conditions has revealed groups of co-regulated metabolites (Weckwerth and Fiehn, 2002). Metabolite levels are thus controlled by both genetic and environmental factors.

The question arises as to whether metabolite concentrations can be considered as quantitative traits in a search for metabolite quantitative trait loci (mQTL) that control their abundance, and whether QTL analyses of the concentrations of all intermediates in a given biochemical pathway can

reveal flux-regulating control points, either to all pathway intermediates or to a subset of intermediates only. Importantly, and in contrast to other complex traits, such as morphological or physiological traits, the molecular structure of the analyzed metabolites and the knowledge of the pathway architecture may already suggest the function of the gene underlying the mQTL. Hence, when the genome sequence of the studied organism is available, the identification of candidate genes with the predicted functions should be possible.

Poplar (*Populus* sp.) has become the model of choice for molecular genetic research on trees (Boerjan, 2005). Genetic maps have been created (Cervera *et al.*, 2001, 2004; Yin *et al.*, 2004), and sequencing of the *Populus* genome has recently been completed (http://www.ornl.gov/ipgc/ and http://genome.jgi-psf.org/Poptr1/Poptr1.home.html). The

genus *Populus*, which consists of approximately 30 species (Cervera *et al.*, 2005), presents a rich diversity of flavonoids in the young leaves, buds and bud exudates, the composition of which is characteristic for each species (Greenaway *et al.*, 1992). In these apical tissues, flavonoids are thought to function both as sunscreens and as defense compounds (Christensen *et al.*, 1998; Dixon *et al.*, 2002). Because the overall structure of the flavonoid pathway is well known (Figure 1) and different flavonoids have distinct UV/visible absorption characteristics (Markham and Mabry, 1975), flavonoid biosynthesis is an ideal model to evaluate the feasibility of detecting mQTL controlling the metabolite levels in a pathway. QTL analyses of the concentrations of closely related metabolites, belonging to a complex, multi-branched pathway, have not been performed in any plant species to date.

In this pilot study, a QTL analysis was carried out on the concentrations of the major flavonoids present in apical tissues of two F$_1$ mapping families of poplar that share a common female parent. QTL analyses of 15 shared flavonoids revealed four robust mQTL that control flavonoid levels, two on linkage group (LG) XIII, one on LG III, and one on LG IV. The chemical structure of the flavonoids, coupled with current knowledge of the pathway architecture and *in silico* mapping of candidate genes, allowed the tentative assignment of a function to three of these mQTL: the mQTL on LG III might be involved in the committed step to flavonoid biosynthesis, i.e. chalcone synthase (CHS), and the two mQTL on LG XIII might act at branch points within the pathway, namely acetylation of the 3-*O* position and methylation of the 7-*O* position in the production of pinobanksin 3-acetate and pinostrobin respectively.

Results

HPLC metabolite profiles

To determine whether metabolite profiles were inherited from parents to offspring, the aromatic compounds present in apical tissues of the F$_1$ families 001 and 002 (Experimental procedures), and of their parents, *Populus deltoides* cv. S9-2, *P. nigra* cv. Ghoy and *P. trichocarpa* cv. V24, were analyzed by reverse-phase HPLC. The UV/visible absorption spectra of the chromatogram peaks suggested the presence of simple phenolics and benzoic acid derivatives as well as

phenylpropanoids and flavonoids. The metabolite profiles were different for the three parents, both qualitatively and quantitatively. Characteristic chromatogram peaks of each parent could be traced back in the chromatograms of their respective progeny, indicating the inheritance of parent-specific compounds in the offspring. Based on analysis of the 15 most abundant chromatogram peaks, the mean broad-sense heritability was shown to vary between 0.55 and 0.82, depending on the quantification method used (Experimental procedures).

Because flavonoids are abundantly present in apical tissues of *Populus* (Greenaway *et al.*, 1992) and have characteristic UV/visible absorption spectra, and because the structure of the flavonoid biosynthetic pathway is well described, this pathway lends itself as an excellent model to evaluate the feasibility of genetical metabolomics of a complex biosynthetic pathway. A total of 29 flavonoids could be clearly distinguished in all individuals of family 001. The chromatograms of family 002 revealed 39 flavonoids. Thirteen of these 39 flavonoids were undetectable in 25–50% of the family 002 individuals, and for some of them chi-squared tests hinted at a 1:1 or 1:3 Mendelian segregation. However, no significant mQTL could be detected for any of the 13 compounds using single-trait QTL analysis of population 87002, suggesting that these flavonoids were below the detection limit rather than absent in part of the family.

Of the 26 flavonoid peaks that were present in all individuals of family 002, spiking indicated that 15 of them were also found in family 001. Because of the restricted number of traits that are accepted by MultiQTL (see below), and to be able to compare the results for both families 001 and 002 that share the common female parent *P. deltoides* cv S9-2, all subsequent analyses were focused on these 15 common peaks (Figure 2).

Flavonoid concentration distributions

The concentration distributions of all 15 flavonoids present in both families were unimodal and in most cases skewed to the right. Most of these flavonoids were clearly present in the common *P. deltoides* cv. S9-2 parent, as expected, but often not detected in the *P. nigra* cv. Ghoy or in *P. trichocarpa* cv. V24 parents (Table 1). The concentrations of some of the compounds were higher in the hybrids than in either parent. This was apparent for flavanone **5** in both families

Figure 1. Flavonoid biosynthesis.
4CL, 4-coumarate:CoA ligase; C3H, *p*-coumarate 3-hydroxylase; C4H, cinnamate 4-hydroxylase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; F3'H, flavonoid 3'-hydroxylase; F4'H, flavonoid 4'-hydroxylase; F3OGlcT, flavonol 3-*O*-glucosyltransferase; F3OMT, flavonol 3-*O*-methyltransferase; FRT, flavonol 3-*O*-glucoside L-rhamnosyltransferase; FS, flavone synthase; FLS, flavonol synthase; HCT, *p*-hydroxycinnamoyl-CoA:shikimate/quinate *p*-hydroxycinnamoyltransferase; AT, acyl-CoA-dependent acyltransferase; OMT, *O*-methyltransferase; PAL, phenylalanine ammonia lyase. Classification of the different flavonoids is based on the nature of the functional group for which the structure is shown in red. The numbers and, if known, structures of the 15 flavonoids used for QTL analysis are shown in bold. Background colors of the flavanones and dihydroflavonols, and of the chalcones, flavones and flavonols are green and red respectively. ?, enzymatic conversion not yet demonstrated; CA, enzymatic conversion suggested by correlation analysis; QTL, enzymatic conversion suggested by quantitative trait locus analysis.
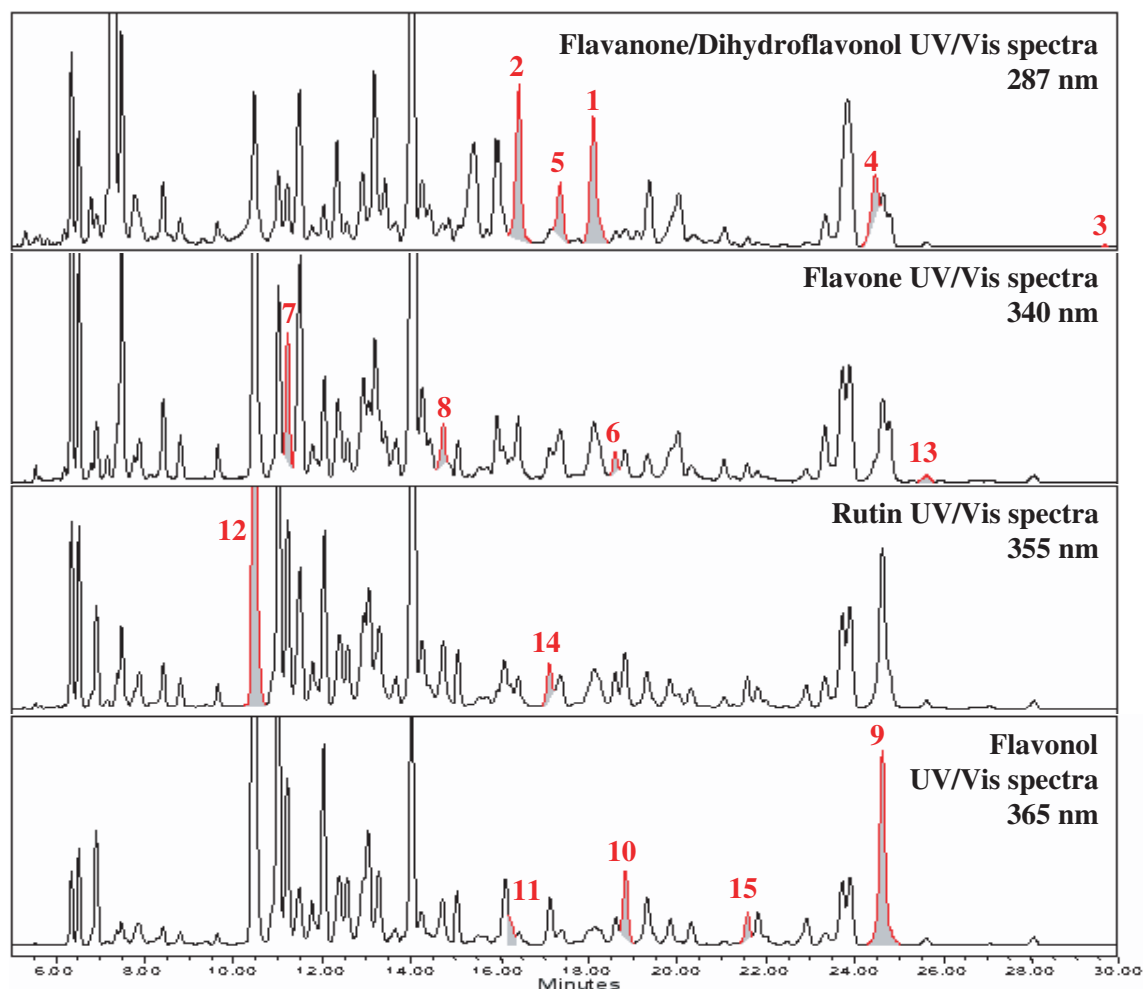
Figure 2. HPLC separation of one individual of population 87001.
Single-wavelength chromatograms were taken at 287, 340, 355 and 365 nm to integrate the flavonoids with a UV/visible absorption spectrum similar to that of a flavanone/dihydroflavonol, a flavone, rutin and a flavonol respectively. The 15 flavonoids in common between families 001 (populations 87001 and 95001) and 002 (populations 87002 and 95002), indicated at their respective wavelengths, are shown in red.

and for rutin **12** in family 002, and was confirmed for both by the additional investigation of nine ramets of each parent and four different cultivars of each of the three poplar species. This phenomenon, called chemical over-expression, is thought to result from the obstruction or elaboration of a pathway in the F$_1$ hybrids, leading to the accumulation of intermediates or (new) end products respectively (Orians, 2000).

Correlation analysis

Because the 15 flavonoids are synthesized from the same biosynthetic pathway (Figure 1), their levels were expected to be highly correlated. Analysis of their correlation coefficients may suggest groups of flavonoids whose levels are co-regulated, and, additionally, reaction steps for which mQTL may be found. To investigate which of the 15 flavonoid concentrations were correlated, the 105 possible correlation

coefficients of the 15 flavonoids were calculated based on the peak height/dry weight (PH/DW) (Table S1). No negative correlations were found. Correlation networks were subsequently generated for the highly correlated ($r > 0.80$) flavonoid concentrations in each population. Figure 3 shows that both general and family-specific associations were evident.

For the two families 001 and 002, the correlation networks showed a strong association between the levels of quercetin **11** and quercetin 3-methyl ether **14** (Figure 3), the substrate and product of flavonol 3-*O*-methyltransferase (F3OMT) respectively. The same enzyme converts galangin **9** to galangin 3-methyl ether **13**, whose abundances were also highly correlated in all populations except 95002.

In all populations, except 95001, eriodictyol **2**, galangin **9**, kaempferol **10** and the unknown flavonol **15** were mutually highly correlated (Figure 3). This result indicates a strong association between the flavanone/dihydroflavonol and flavonol branches of flavonoid biosynthesis,

Table 1 Descriptive statistics of flavonoid levels

Name	Class	Compound (ng mg^{-1} dry weight)			87001 95001	87002 95002
		P.d.	*P.n.*	*P.t*		
Pinobanksin	Dihydroflavonol (**1**)	615	ND	31.1	761 (400) 1070 (700)	68.7 (69) 14.5 (17)
Eriodictyol	Flavanone (**2**)	1120	11.8	ND	697 (370) 917 (610)	193 (170) 55.2 (64)
Pinostrobin	Flavanone (**3**)	2.04	ND	ND	2.97 (2.1) 4.91 (3.8)	0.342 (0.39) 0.289 (0.31)
Pinobanksin 3-acetate	Dihydroflavonol (**4**)	673	ND	ND	307 (210) 339 (570)	95.4 (87) 36.5 (41)
Unknown	Flavanone* (**5**)	ND	ND	ND	205 (110) 142 (120)	237 (210) 78.1 (118)
Apigenin	Flavone (**6**)	12.6	29.3	ND	63.4 (37) 73.8 (612)	59.3 (59) 3.7 (33)
Unknown	Flavone (**7**)	133	ND	26.3	175 (170) 141 (120)	257 (250) 83.0 (93)
Unknown	Flavone (**8**)	103	ND	ND	65.7 (47) 29.4 (30)	17.4 (25) 10.5 (22)
Galangin	Flavonol (**9**)	253	ND	ND	315 (170) 473 (320)	62.2 (73) 22.7 (25)
Kaempherol	Flavonol (**10**)	227	14.0	0.752	142 (90) 207 (140)	225 (210) 65.7 (79)
Quercetin	Flavonol (**11**)	107	18.8	10.8	131 (100) 211 (160)	198 (170) 60.9 (91)
Rutin	Flavonol (**12**)	ND	2440	ND	841 (550) 837 (590)	141 (130) 88.4 (93)
Galangin 3-methyl ether	Flavonol (**13**)	12.0	ND	ND	16.5 (8) 20.5 (12)	2.76 (2.4) 1.48 (5.5)
Quercetin 3-methyl ether	Flavonol (**14**)	172	3.42	ND	658 (460) 75.0 (58)	46.9 (39) 38.5 (53)
Unknown	Flavonol (**15**)	102	1.22	ND	54.8 (31) 67.6 (47)	54.3 (45) 16.1 (17)

Means and standard deviations (SD) are given for the concentrations (ng mg^{-1} dry weight) of the 15 common flavonoids in families 001 (populations 87001 and 95001) and 002 (populations 87002 and 95002), together with the parental concentration values (*P.d.*, *P.n.* and *P.t.*). For description of the families, see Experimental procedures. ND, not detected; *P.d.*, *Populus deltoides* cv. S9-2; *P.n.*, *P. nigra* cv. Ghoy; *P.t.*, *P. trichocarpa* cv. V24. *UV/visible spectrum was similar to that of a flavanone or dihydroflavonol.

represented by eriodictyol **2** and by galangin **9**, kaempherol **10** and the unknown flavonol **15** respectively. Notably, the correlation networks did not show a strong correlation between the levels of any of the flavones, i.e. apigenin **6** and the unknown flavones **7** and **8**, and the levels of either flavanones/dihydroflavonols or flavonols. Also, the two flavanones, pinostrobin **3** and the unknown flavanone **5**, were not consistently correlated with the level of any other flavonoid.

In addition to general associations, family-specific associations between flavanone/dihydroflavonol and flavonol biosynthesis also prevailed in the correlation networks. In family 001, the concentrations of pinobanksin **1**, eriodictyol **2** and galangin **9** were highly correlated, whereas strong correlations between the levels of eriodictyol **2**, pinobanksin 3-acetate **4**, kaempherol **10**, quercetin **11**, quercetin 3-methyl ether **14** and the unknown flavonoid **15** were prominent in family 002 (Figure 3). Family 002 was further characterized

by a high correlation between the levels of the flavone **7** and rutin **12** (Figure 3).

Taken together, both general and family-specific correlations were found. Within each family, most correlations were consistently observed in both populations. A closer examination of the correlation networks in each population did not reveal groups of flavonoids of a given class, i.e. no mutually highly correlated clusters were found that contained all flavones, all flavonols or all flavanones. In contrast, both general and family-specific correlations pointed to a tightly associated biosynthesis of specific flavanones/ dihydroflavonols and flavonols.

QTL analysis of flavonoid concentrations

To reveal loci that control the flux within flavonoid metabolism, mQTL were searched for the different flavonoid concentrations. A multi-trait approach was applied by using
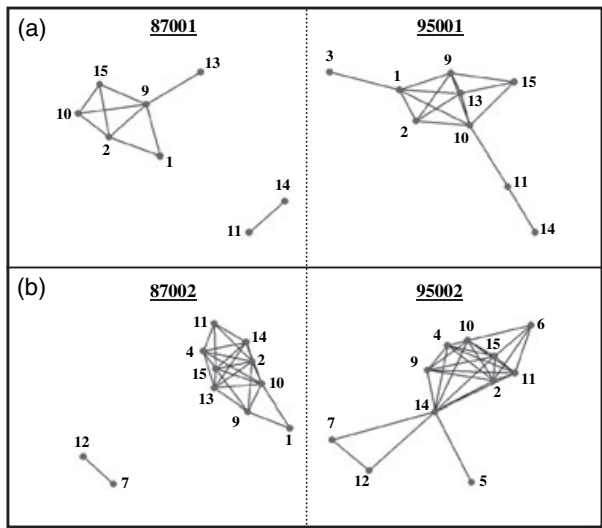
Figure 3. Metabolite correlation networks.
Correlation networks, obtained by applying the Fruchterman–Reingold 2D algorithm, are shown for each population and were generated by using the levels of the 15 flavonoids in common between families 001 and 002 and applying a threshold correlation coefficient ($r > 0.8$). Vertices and edges represent flavonoids and strong correlations respectively. The Pearson product–moment correlation coefficients are listed in Table S1.

maximum-likelihood interval mapping because of the multiple traits and the high correlations that were often found between the different flavonoids. However, the higher the number of traits in a multi-trait approach, the higher the probability that multiple loci along the chromosome affect the multivariate trait and the higher the chance of detecting so-called 'ghost' QTL caused by the interfering effect of
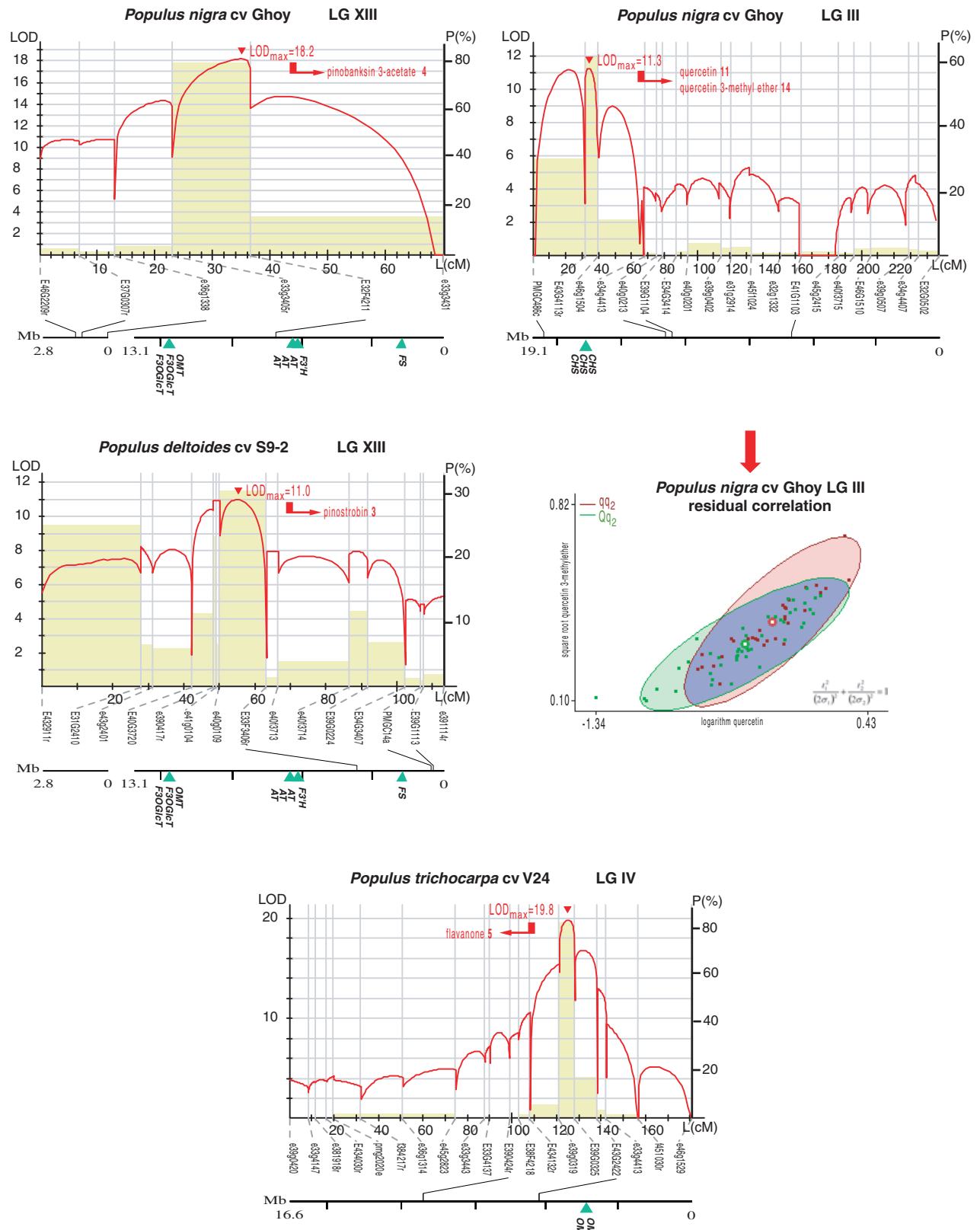
linked loci (Jiang and Zeng, 1995; Knott and Haley, 2000; Korol *et al.*, 2001; Martínez and Curnow, 1992). Therefore, single-trait QTL analysis with both regression and a non-parametric Wilcoxon test was performed as an alternative and complementary approach. Furthermore, because ratios of compound concentrations are more robust than individual metabolite levels (Fiehn, 2003; Morreel *et al.*, 2004; Steuer *et al.*, 2003), we calculated the 105 possible ratios between the peak heights of all 15 flavonoids of the families 87001 and 87002, logarithmically transformed them to so-called log ratios (Birks and Kanowski, 1993) and used them for univariate or single-trait QTL analysis. This strategy increases the chance of detecting mQTL that control the differential synthesis of two intermediates present in the same pathway. The QTL results obtained by the different methods (mIM and single-trait QTL analyses of log ratios) are presented in Table 2, Figure 4 and Tables S2 and S3. From these data, robust mQTL were assigned based on the criteria explained in the Experimental procedures. The mQTL of flavonoid concentration levels that were obtained for populations 87001 and 87002 are given below.

In family 001, multi-trait interval mapping (mIM) analysis revealed two mQTL on the genetic map of *P. nigra* cv. Ghoy (hereafter designated *P.n.* map), on LG XIII and on LG III, and one mQTL on the map of *P. deltoides* cv. S9-2 (designated *P.d.* map), on LG XIII. The highest LOD score (18.2) was observed at marker E32F4211 on LG XIII of the *P.n.* map (Figure 4, likelihood map in red). Bootstrapping results (Figure 4, yellow bars) indicated an almost 80% chance that the mQTL occurred in the 13 cM marker interval e33g3405r–E32F4211. Examination of the log ratios that were affected by the mQTL (Table 2) revealed the

Table 2 Quantitative trait locus-associated LOD scores and flavonoid ratios affected by the QTL

Genetic map	LG	LOD (mIM)		Ratios (univariate)
Family 001		87001	95001	87001
P. d.	XIII	11.0[a]	15.2[a]	1/3[a], 2/3[a], 4/3[b], 5/3[b], 6/3[a], 9/3[a], 10/3[a], 11/3[a], 12/3[a], 13/3[a], 14/3[a], 15/3[a]
	XV			13/10[b]
P.n.	III	11.3[b]	11.9[a]	1/11[a], 2/11[a], 6/11[a], 9/11[a], 10/11[a], 13/11[a], 15/11[a], 6/14[a], 9/14[b], 13/14[a], 15/14[a]
	XIII	18.2[a]	9.5[b]	1/4[a], 2/4[a], 6/4[a], 9/4[a], 10/4[a], 13/4[a], 15/4[a]
Family 002		87002	95002	87002
P.d.	XIII	11.0[a]		
P.t.	IV	19.8[a]	12.1[a]	1/5[a], 2/5[a], 3/5[a], 4/5[a], 7/5[a], 9/5[a], 10/5[a], 12/5[a], 13/5[a], 14/5[b], 15/5[a], 12/7[a]
	V	13.8[a]		8/6[b], 15/6[a]
	XIII		10.4[a]	

LOD scores obtained by multi-trait interval mapping (mIM) are given for families 001 and 002 (see Experimental procedures). Additionally, in populations 87001 and 87002, the 105 possible log ratios between the 15 different flavonoids (numbers refer to the compounds listed in Table 1) were logarithmically transformed and subjected to univariate QTL analysis (see Experimental procedures). Based on *post hoc* permutations, genome-wise significance values below 0.05 and below 0.10 are indicated by the superscripts a and b respectively. LG, Linkage group.

importance of this locus on the abundance of pinobanksin 3-acetate **4** (Table S3). In agreement, the highest value (24%) for the variance explained by the mQTL, as determined by mIM, was associated with pinobanksin 3-acetate **4** (Table S2).

A second mQTL on the *P.n.* map was located on LG III and reached its maximum (LOD 11.3) in the interval E43G4113r–e46g1504. The mQTL had a 60% probability of occurrence in this 10 cM interval based on bootstrapping results (Figure 4, yellow bars). Although the highest values for the variance explained by the mQTL (Table S2) were associated with quercetin **11** and quercetin 3-methyl ether **14** in population 87001 only, all single-trait mQTL for log ratios involving either quercetin **11** or quercetin 3-methyl ether **14** co-localized with the mQTL predicted by mIM (Table S3). Therefore, we concluded that the mQTL is involved in the biosynthesis of the latter two compounds.

As described above, the levels of quercetin **11** and quercetin 3-methyl ether **14** were strongly correlated in all populations. In the case of population 87001, both compounds were isolated from the remainder of the correlation network because the abundance of neither one strongly correlated with any of the other 13 flavonoid levels (Figure 3). If the mQTL on LG III, obtained by mIM, were indeed implicated in the strong association between the levels of both quercetin **13** and quercetin 3-methyl ether **14** as suggested by the single-trait mQTL analyses of the log ratios, it would explain a major part of their concentration co-variance. Indeed, the transformed concentration levels of quercetin **11** and quercetin 3-methyl ether **14** had an initial correlation coefficient of 0.89 in population 87001, whereas the correlation was reduced to 0.58 when the mQTL effect was eliminated (Figure 4). mIM indicated that 19% and 14% of the concentration variances of quercetin **11** and quercetin 3-methyl ether **14** were explained by this mQTL respectively. Taken together, these data point to an mQTL on LG III that affects the concentrations of both quercetin **11** and quercetin 3-methyl ether **14**.

The mQTL on LG XIII of the *P.d.* map was specifically associated with the abundance of pinostrobin **3**, and the maximum LOD score (11.0) was found in the interval e40g0109–E33F3406r (Figure 4). Approximately 30% of the variance in pinostrobin **3** concentrations was explained by this mQTL.

In family 002, four mQTL were detected using mIM, of which one was in both populations of this family, i.e. in 87002 and 95002 (Table S2). This mQTL was located on LG IV of the genetic map of *P. trichocarpa* cv. V24 (*P.t.* map), between the markers e39g0319 and E39G0325 (Figure 4). Bootstrapping results indicated a >80% probability that the mQTL was located in this interval, which is less than 10 cM. An LOD score of 19.8 was obtained, and the mQTL affected significantly only the concentration of the flavanone **5** when the QTL results of the log ratios were surveyed (Table 2), explaining approximately 44% of its concentration variance (Table S2).

Notably, all mQTL were involved in the biosynthesis of one or two flavonoids that were only moderately correlated to all other flavonoid levels as revealed by the correlation networks (Figure 3). Furthermore, no mQTL were detected for the total peak height relative to the dry weight taken as an estimate of the total amount of aromatics.

Identification of candidate genes

Homologues of all known flavonoid biosynthesis genes were searched for in the poplar genome to identify possible candidate genes for the detected mQTL. The number of homologues found for each structural gene is shown in Table S4. For chalcone isomerase (*CHI*), flavanone 3-hydroxylase (*F3H*) and flavonoid 3′-hydroxylase (*F3′H*), only one homologue was detected in the poplar genome, i.e. on LG X, LG V and LG XIII respectively. Two to five homologues were found for the remaining flavonoid biosynthesis genes, and they were distributed over the genome.

To reveal whether some of the flavonoid biosynthesis (Figure 1) gene homologues were present in the mQTL, amplified fragment length polymorphism (AFLP) markers in the mQTL regions were sequenced and mapped *in silico* on the poplar genome sequence (Figure 4). Interestingly, LG III contained three *CHS* homologues that were located in the 90% confidence interval of the mQTL, namely between markers PMGC486c and e40g0213. In addition, a *F3′H*, a flavone synthase (*FS*) and two flavonol 3-*O*-glucosyltransferase (*F3OGlcT*) homologues were found on LG XIII. No homologues of flavonoid biosynthesis genes were detected on LG IV (Figure 4).

Because we hypothesize that the mQTL on LG XIII controlling the abundance of pinostrobin **3** and pinobanksin

Figure 4. Metabolite quantitative trait loci position.

The presented mQTL were obtained by mIM with the peak height/dry weight ($\mu V\ mg^{-1}$) of the 15 flavonoids in common between families 001 and 002. Map positions (cM) and marker names (Cervera *et al.*, 2004) are given below each LG. The QTL likelihood map and the probability that the mQTL is present in each marker interval as determined by bootstrapping are shown in red and yellow respectively. Scales presenting the LOD score and the probability of occurrence in a marker interval are on the left and right side of each QTL likelihood map respectively. The position of the maximum LOD score is indicated by a triangle. In each QTL likelihood profile, the flavonoid(s) whose concentration(s) is (are) affected by the mQTL is (are) noted. The horizontal line below the QTL likelihood profile represents the physical length in megabases (Mb) of the assembled LG. Positions of candidate genes are indicated by green triangles, and the positions of the genetic markers by thin lines. In addition, the bivariate standard normal distribution of quercetin **11** and quercetin 3-methyl ether **14** is presented for each genotype group of the mQTL at LG III of the *Populus nigra* genetic map. Concentration values of quercetin **11** were logarithmically transformed, and those of quercetin 3-methyl ether **14** by taking the square root, and are given on the abscissa and ordinate respectively. *AT*, acyl-CoA-dependent acyltransferase; *CHS*, chalcone synthase; *F3′H*, flavonoid 3′-hydroxylase; *F3OGlcT*, flavonol 3-*O*-glucosyltransferase; *FS*, flavone synthase; *OMT*, *O*-methyltransferase.

3-acetate **4** are a methyltransferase and an acetyltransferase (see Discussion), hidden Markov marker (HMM) profiles of *O*-, *C*-, *N*- and *S*-methyltransferases and of CoA-dependent *O*-acyltransferases were constructed and used to search for homologues on the poplar genome (see Experimental procedures). One *O*-methyltransferase homologue and three CoA-dependent *O*-acyltransferases were mapped to the LG XIII (assembly version 1.0, June 2004; Figure 4).

Discussion

Most flavonoid levels are highly correlated

Reverse-phase HPLC profiles of methanol-soluble phenolics from apical tissues or mature leaves of the poplar F_1 families and their parents indicated that parent-specific peaks are inherited and that a significant heritability can be calculated for most peaks. In the HPLC chromatograms of the apical tissue extracts, 29 and 39 compounds had UV/visible absorption spectra of flavonoids in families 001 and 002, respectively, of which 15 peaks were found in both families.

The concentrations of these 15 common flavonoids in both families 001 and 002 were highly positively correlated. This observation is in contrast with most metabolomics studies in which only a small number of metabolite pairs were highly correlated and the majority poorly, even when the metabolite pairs belong to the same pathway (Camacho *et al.*, 2005). However, the latter metabolite correlation networks were constructed only for individuals of the same genotype (Fiehn, 2003; Weckwerth *et al.*, 2004), whereas our correlation networks are based on metabolite abundances measured in a segregating population, thus including a component of genetic variation.

In both families, a strong correlation between the abundances of neighboring pathway intermediates was noticed between quercetin **11** and quercetin 3-methyl ether **14** levels as well as between galangin **9** and galangin 3-methyl ether **13** levels. These results indicate ready production of the 3-methyl ethers from their respective flavonols despite the endothermic nature of the reaction, and hence suggest that F3OMT is not rate-limiting in the production of quercetin 3-methyl ether **14** and galangin 3-methyl ether **13**, irrespective of the genetic background.

Additionally, both general and family-specific clusters of mutually correlated levels of flavonoids belonging to distinct parts of the pathway were present. These observations are in accordance with simulation studies of metabolic correlation networks that often reveal strong correlations between seemingly distant metabolites (Steuer *et al.*, 2003). Such strong correlations between distant pathway intermediates may arise from short-term regulation, such as metabolic feedback control, or long-term regulation, such as transcriptional activation, or from compartmentalization or the presence of metabolons (Winkel, 2004). Here, strong corre-

lations were mainly found between flavanone/dihydroflavonol and flavonol levels.

QTL are associated with the flux-controlling steps in a pathway

Although the determination of flux-controlling steps within a pathway from correlation networks is complicated by the presence of genetic variation (Camacho *et al.*, 2005), the latter type of variation allows the detection of QTL. By assuming that one gene is responsible for the observed QTL effect on the abundance of a certain compound and following the theory of metabolic control analysis (Heinrich and Rapoport, 1974; Kacser and Burns, 1973, 1981) and simulation studies (Bost *et al.*, 1999), the gene product behind the QTL may be considered to exert a rate-limiting effect in half of the population and might be regarded as a control point in the biosynthetic flux towards that compound. Because of the flux-controlling effect of QTL, searching for mQTL might be an approach to obtain fluxome data from metabolome studies.

In this study, all mQTL affected the levels of one or two flavonoids that were low or only moderately correlated to all other flavonoids. This suggests that a correlation analysis prior to the QTL analysis may already indicate for which biochemical conversions mQTL can be detected. Such an approach would reduce the number of QTL analyses to be executed, making the QTL analyses in a metabolome-wide study involving thousands of peaks much more amenable.

QTL analysis of flavonoids

We have chosen a multivariate QTL analysis because it performs better than single-trait analyses in cases of functionally related traits, such as the concentration of pathway intermediates. Additionally, single-trait analyses of all mutual log ratios between the flavonoids were performed. Six mQTL were placed on the parental maps of populations 87001 and 87002 by means of mIM, five of which were also detected by single-trait QTL analysis. In a second experiment involving populations 95001 and 95002, five mQTL were found, four of which were also detected for populations 87001 and 87002 with mIM as well as single-trait analyses. Because these mQTL affected log ratios of flavonoid levels, they are expected to function within flavonoid metabolism. These four robust mQTL are discussed below, with the assumption that the mQTL effect is not a composite effect of different linked loci.

An mQTL on LG XIII of P. nigra *is involved in the production of pinobanksin 3-acetate* **4**

Both multi-trait and single-trait analyses indicated that the abundance of pinobanksin 3-acetate **4** was affected by the

mQTL on LG XIII of the *P.n.* map. We speculate that this mQTL, which explained 24% of the variation in the concentration of pinobanksin 3-acetate **4**, controls the 3-*O*-acetylation step in the biosynthesis of pinobanksin 3-acetate **4** from pinobanksin **1** (Figure 1), because this mQTL affected specifically all ratios with pinobanksin 3-acetate **4**, including the ratio between pinobanksin **1** and pinobanksin 3-acetate **4**. The few known *O*-acetyltransferases that are operative in secondary metabolism belong to the acyl-CoA-dependent BAHD superfamily (EC 2.3.1.–) (Ma *et al.*, 2005), whose acronym is derived from the first four enzymes of this family isolated from plants, i.e. benzyl alcohol acetyl, anthocyanin-*O*-hydroxycinnamoyl, anthranilate-*N*-hydroxycinnamoyl/benzoyl and deacetylvindoline acetyltransferase. In accordance with our hypothesis that the mQTL might control the 3-*O*-acetylation of pinobanksin **1**, three members of the BAHD superfamily were found near the maximum LOD score on LG XIII. The three genes were located in tandem and their corresponding proteins shared 78–89% identity and 89–95% similarity. The closest *Arabidopsis thaliana* homologues (58–60% identity and 70–74% similarity) are two genes with unknown function (At5g17540 and At3g03480).

The production of quercetin 11 and its 3-methyl ether 14 is co-regulated by an mQTL on LG III of P. nigra

An mQTL on LG III of the *P.n.* map explained 19% and 14% of the variation in the concentrations of quercetin **11** and quercetin 3-methyl ether **14** respectively. Both correlation and QTL analysis revealed a prominent contribution of this mQTL to the association between both compounds. Because none of the levels of the other flavonoids, especially the precursors eriodictyol **2** and kaempferol **10**, are affected by this mQTL, channeling might play a role in the production of quercetin **11** and quercetin 3-methyl ether **14** (Figure 1).

Indeed, several studies have supported the existence of an enzyme complex between CHS and CHI (Winkel, 2004), and co-immunoprecipitation studies in Arabidopsis have shown that the two enzymes may interact with F3H (Burbulis and Winkel-Shirley, 1999). Both F3H and FLS are 2-ketoglutarate-dependent dioxygenases. The latter enzyme has been shown to be bifunctional in *Citrus unshiu*, catalyzing the 2-hydroxylation of (2S)-naringenin to aromadendrin and the further oxidation to kaempferol **10** (Lukačin *et al.*, 2003). Hence, some of these 2-ketoglutarate-dependent dioxygenases appear to be involved in multiple flavonoid biosynthesis steps. Also the interaction of CHS and CHI with F3′H has been suggested by the observation that CHS and CHI co-localized in wild-type Arabidopsis roots but not in those of *transparent testa7* mutants. The latter mutants are defective in F3′H, a cytochrome P450-dependent monooxygenase that is thought to act as a membrane anchor (Saslowsky and Winkel-Shirley, 2001). Thus, the production of quercetin **11** and quercetin 3-methyl ether **14** may proceed from *p*-coumaroyl-CoA or caffeoyl-CoA by a metabolic channel involving the CHS/CHI/(F3H/)(F3′H/)FLS reactions. Given that three *CHS* homologues were found near the maximum LOD score position on LG III, we speculate that CHS controls the flux through this metabolon.

Pinostrobin 3 biosynthesis is controlled by a locus on LG XIII of P. deltoides

The mQTL on LG XIII of the *P.d.* map controls specifically the concentration of pinostrobin **3** in family 001. Thus, this locus may be associated with the 7-*O*-methylation of pinocembrin (Figure 1). Bioinformatics revealed only one *O*-methyltransferase (*OMT*) homologue on the LG XIII, although in the QTL regions, sequence gaps may still exist in the currently assembled poplar genome (Version 1.0, June 2004). Blasting of this OMT against the database of the National Center for Biotechnology Information showed only modest similarity to experimentally proven OMTs. The closest OMTs for which a catalytic activity has been shown is a caffeic acid *O*-methyltransferase (COMT) from *Rosa chinensis* (Wu *et al.*, 2003), which shared 58% identity and 74% similarity. Other close homologues were an (*R,S*)-reticuline 7-*O*-methyltransferase from *Papaver somniferum* (identity 49%; similarity 67%; Ounaroon *et al.*, 2003) and a hydroxycinnamic acid/hydroxycinnamoyl-CoA OMT (AEOMT) from *Pinus taeda* (identity 49%; similarity 68%; Li *et al.*, 1997). Hence, no function has yet been described for the *OMT* gene on LG XIII. Further experiments are needed to evaluate whether the corresponding protein is able to catalyze the methylation of pinocembrin to pinostrobin **3**, although this experiment may not be entirely conclusive because OMTs can typically methylate a variety of substrates *in vitro*, sometimes belonging to different structural classes (Ibdah *et al.*, 2003).

An mQTL located on LG IV of P. trichocarpa affects the unknown flavanone 5 level

In family 002, an mQTL was found on LG IV of the *P.t.* map that controlled the level of an unknown flavanone **5**, which is a compound that could not be detected in either of the parents, i.e. *P. deltoides* cv S9-2 or *P. trichocarpa* cv. V24, but for which the most significant mQTL with the smallest confidence interval was detected. The chemical over-production of this flavanone **5** remains unexplained, but might be caused, for instance, by the combination, in the F$_1$ hybrids, of two different enzyme functions acting on a common substrate. Although it was not possible to elucidate the structure of this flavanone **5**, it is probably derived from pinocembrin, naringenin or eriodictyol **2**, the three flavanones of which all the other flavanones are derived. We hypothesize that the mQTL affects this derivation step.

Conclusions

Metabolite profiling of flavonoids in poplar was chosen as a model to scan for mQTL that control the levels of these metabolites. Based on the quantification method, the identities of the metabolites affected by the mQTL, the architecture of flavonoid biosynthesis and the correlation analysis, tentative mQTL functions could be assigned. Some mQTL might control the flux at key enzymatic steps, as hypothesized for the CHS-associated mQTL on LG III, whereas others might control the flux at particular pathway branches, such as the putative 7-*O*-methyltransferase- and 3-*O*-acetyltransferase-associated mQTL detected on LG XIII. All mQTL were involved in the biosynthesis of one or two flavonoids that were only moderately correlated to all other flavonoid levels. Our data show that metabolite profiles can be used for QTL analysis to reveal loci that control the flux through a complex, multi-branched pathway of related molecules, such as flavonoid biosynthesis. To our knowledge, we demonstrate for the first time the potential of genetical metabolomics.

Experimental procedures

Pedigrees and genetic maps

The poplar F_1 family 001 (152 individuals) was derived from controlled crosses between *Populus deltoides* cv. S9-2 (female) and *P. nigra* cv. Ghoy (male) performed in 1987 (population 87001, 88 individuals) and in 1995 (population 95001, 64 individuals). The F_1 family 002 (202 individuals) was derived from crosses between the same *P. deltoides* cv. S9-2 (female) and *P. trichocarpa* cv. V24 (male) in 1987 (population 87002, 78 individuals) and in 1995 (population 95002, 124 individuals). Construction of the AFLP linkage maps of these three parents using populations 87001 and 87002 and their alignment with microsatellites have been described by Cervera *et al.* (2001).

Metabolite profiling

Woody cuttings (one cutting per genotype) of F_1 populations 87001, 95001, 87002 and 95002 and their parents were planted in the greenhouse and grown for 4 months (from February to June) until a plant height of approximately 2 m. Because the apical tissues are the richest in flavonoid diversity and to minimize developmental and environmental variations between organs of different ages, the apical part of each plant, consisting of the three youngest internodia and leaves (length of largest leaf blade approximately 2.5 cm, fresh weight approximately 300 mg), was harvested when photosynthesis had reached a steady state between 14:00 and 15:00 hours (Lohaus *et al.*, 1995; Riens *et al.*, 1991; Winter *et al.*, 1993). After the plant material had been ground in a mortar with liquid nitrogen, it was extracted in 15 ml methanol:isopropanol:tetrahydrofuran (70:15:15; v:v:v) and 0.5 ml of the supernatant was lyophilized. The residue was dissolved in 1.6 ml 0.1% aqueous trifluoroacetic acid (TFA) and cyclohexane (50/50; v/v). The remaining residue of cell debris of each sample was freeze-dried and weighed (dry weight approximately 70 mg). The aqueous phase was analyzed on a C18 reverse-phase HPLC column (Luna C18(2), 5 μm, 4.6 mm × 250 mm; Phenomenex, Torrance, CA, USA) with a flow rate of 1.5 ml min^{-1} and a column temperature of 40°C. A 625 LC pump (Waters, Milford, MA, USA) changed the mobile phase gradually from 83% buffer A [100% MilliQ water (Millipore, Bedford, MD, USA), 1% acetonitrile, 0.1% TFA] to 77% buffer B (75% acetonitrile, 25% methanol, 1% MilliQ water, 0.1% TFA) within 29 min. UV/visible spectra between 250 and 450 nm were taken on a 996 PDA detector (Waters).

Quantification of flavonoids

The flavonoids of interest were quantified by using the maxima in their UV/visible absorption spectra: 287 nm for flavanones/dihydroflavonols, 340 nm for the flavones, 355 nm for flavonoids with a rutin **12**-like UV/visible spectrum, and 365 nm for flavonols. Single-wavelength chromatograms were integrated using Millenium[32] software (Waters). The flavonoid abundance was obtained by standardizing the peak height to the dry weight (PH/DW). This HPLC method yielded coefficients of variation of 8.2%, 8.6% and 9.0% for the abundance of kaempferol, pinobanksin and apigenin, respectively, based on the weekly repetitive separation of a biological sample for a period of 6 weeks. For the descriptive statistics (Table 1), external calibration was performed using dilution series of flavonoid standards. The calculated slope of the UV/visible absorbance response to the amount of standard obtained for naringenin, apigenin and kaempferol was used to express the concentrations of the flavanones/dihydroflavonols, flavones and flavonols, respectively, as ng mg^{-1} dry weight.

Quantification reproducibility

The reproducibility of the quantification method for methanol-soluble phenolics, in general, was evaluated by correlation of the peak abundances between the two halves of a leaf. A mean correlation coefficient of 0.88 was obtained based on 70 integrated HPLC peaks that were present in both halves of the fifth leaf of ten greenhouse-grown ramets of the same genotype.

Heritability estimation

Ten ramets of individual 87001/63 and one ramet for each of ten randomly chosen individuals of population 87001 were grown. The fifth leaf was extracted and analyzed on HPLC as described above. The heights of the 15 most abundant peaks, of which five corresponded to flavonoids, were determined based on a max-plot between 250 and 450 nm. Quantification was performed with three different standardizations, i.e. by standardizing the peak height: (i) to the sum of the heights of all chromatogram peaks, yielding the percentage of peak height (%PH); (ii) to the fresh weight (PH/FW); and (iii) to the amount of an internal standard (2-guanidinobenzimidazole; Sigma-Aldrich, St Louis, MO, USA) that was added before grinding in an amount proportional to the fresh weight (PH/IS). Based on a one-sided *F* test, significantly less variation ($\alpha = 0.05$) between ramets of the same genotype than between different genotypes was found for the abundance of 13, 12 and six of the 15 peaks when using %PH, PH/FW and PH/IS respectively. The heritability distributions were skewed to the higher values for all three standardization procedures, with means of 0.70, 0.82 and 0.55 respectively.

Characterization of common flavonoids

Fifteen flavonoids (numbered **1** to **15** in Figure 1) were found in common between both F_1 families. LC/MS and/or NMR analyses (Tables S5 and S6, Figure S1), together with the spiking of standards resulted in the identification of eleven of them. Pinobanksin **1** was purchased from Oy ArboNova Ab (Turku, Finland); naringenin, galangin **9** and quercetin **11** were purchased from Sigma-Aldrich; eriodictyol **2**, apigenin **6**, kaempherol **10** and rutin **12** from Roth (Karlsruhe, Germany); pinostrobin **3** from Lancaster Synthesis (Heysham, Lancashire, UK); pinobanksin 3-acetate **4** and quercetin 3-methyl ether **14** from Apin Chemicals (Abingdon, UK). Galangin 3-methyl ether **13** was kindly provided by Guy Lemière, Antwerp University, Belgium (Deng *et al.*, 1997). Based on their UV/visible absorption and/or MS/MS spectra, the remaining four unidentified flavonoids could be further distinguished as a flavanone/dihydroflavonol **5** (λ_{max} = 283 nm), two flavones **7** (λ_{max} = 252 and 347 nm; molecular mass 286 g mol^{-1}) and **8** (λ_{max} = 266 and 336 nm), whereas compound **15** (λ_{max} = 259 and 351 nm) could be either a flavone or a flavonol (Cuyckens and Claeys, 2004; Markham and Mabry, 1975). These four flavonoids were classified into one of the flavonoid subgroups, i.e. as a flavanone/dihydroflavonol, flavone or flavonol.

Verification of chemical over-expression

Ten individuals of families 001 and 002, nine ramets of each of their parents and one ramet of four genotypes of *P. deltoides* (S620-225, S4-52, S620-565 and V5xV1-60), *P. nigra* (Ogy, Oosterzele, Pap4xOgy-56 and Woodecq-3) and *P. trichocarpa* (5618-8, 73022-235, 73021-17 and 73023-5) were planted and grown for three years (field plantation of IBNO at Grimminge, Belgium). The apical tissues of the main stems were harvested and analyzed on HPLC as described above.

Statistical and QTL analysis

Descriptive statistics and Pearson product–moment correlation coefficients were calculated with SPSS 9.0 (SPSS, Chicago, IL, USA). Metabolite correlation networks were generated by the Fruchtermann–Reingold 2D layout algorithm with Pajek software version 1.04 (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

Marker and QTL analyses followed a pseudo-testcross model (Grattapaglia and Sederoff, 1994). To obtain evenly spaced markers in populations 87001 and 87002, accessory markers were used for QTL analysis in addition to framework markers. The segregations of the AFLP markers described in Cervera *et al.* (2001) were also analyzed in populations 95001 and 95002. AFLP markers that could be clearly scored were subsequently used for QTL analysis (Table S7) to verify the mQTL obtained in populations 87001 and 87002.

To account for the multiple traits (i.e. multiple flavonoid concentrations), QTL analysis was based on the one-QTL multi-trait interval mapping (mIM) procedure included in MultiQTL (Korol *et al.*, 2001). MultiQTL performs a canonical transformation procedure, which takes into account the marker interval before the actual maximum likelihood analysis. The power and precision of the parameter estimation in a multi-trait analysis is often higher than that in a single-trait QTL analysis where no advantage of the correlated structure is taken. Furthermore, the statistical power is highest when negative residual correlations between the traits exist or when opposite effects of the QTL on the different traits are present (Jiang and Zeng, 1995; Knott and Haley, 2000; Korol *et al.*, 2001). QTL

analysis was performed with the following settings in MultiQTL: population, backcross; mapping function, Kosambi; starting points for solution, 50. Neither marker restoration nor LOD normalization were applied. Flavonoid concentrations were subjected to Box–Cox transformations prior to QTL analysis to obtain multivariate normality. Chromosome-wise significance levels and 95% confidence intervals were based on 1000 permutations and 100 bootstraps respectively. Means (*m*), QTL effects (*d*) and residual standard deviations (σ_{res}) for the trait values were calculated according to the linear model:

$$Z = m + 0.5dg + \sigma_{res} \tag{1}$$

Z is the individual's trait value and *d* is the effect of substitution aa $\rightarrow$ Aa with respect to the mean trait value. The genotype at locus A/a is denoted by *g* (–1 for aa and 1 for Aa). The variance explained by the QTL was calculated according to a one-way analysis of variance model:

$$\sigma_{exp}^2 = (nd^2 + \sigma_{res}^2)/[nd^2/2 + (n+1)\sigma_{res}^2] \tag{2}$$

where σ_{exp}^2 is the explained variance by the QTL, and *n* is the mean of the number of individuals within the aa and Aa genotype groups. The one-QTL single-trait interval mapping procedure in MultiQTL was employed with the same software settings as mentioned for mIM analysis, for flavonoid levels that were qualitatively distributed and for the total amount of aromatics as determined by the total PH/DW.

In addition to multi-trait analysis, the ratios of the 15 flavonoid levels detected in populations 87001 and 87002 were logarithmically transformed to log ratios (Birks and Kanowski, 1993; Shepherd *et al.*, 1999) and subsequently subjected to single-trait regression and non-parametric Wilcoxon tests, included in the software HSQMV4 (Coppieters *et al.*, 1998). Chromosome-wise significance thresholds and 95% confidence intervals were defined by means of 1000 permutations and 100 bootstrap experiments with the Wilcoxon test.

A Bonferroni correction was applied to obtain a genome-wise significance threshold (α = 0.10 and 0.05). For multi-trait QTL analysis, the Bonferroni correction accounted for the number of LGs, whereas it accounted for the number of LGs as well as for the number of traits in the case of the single-trait Wilcoxon test.

The mQTL of flavonoid concentration levels that were obtained for populations 87001 and 87002 with the different statistical procedures (mIM and single-trait QTL analyses of log ratios) and that were verified by the mIM analysis of populations 95001 and 95002, respectively, were retained. Following the assignment of robust mQTL by using these rigorous QTL selection criteria, the compounds whose concentrations were affected by the mQTL still needed to be determined. This was based on those flavonoids for which: (i) the mIM-determined QTL explained most of the concentration variance in both populations of each family; and (ii) the single-trait mQTL for the flavonoid log ratios, obtained with the populations 87001 and 87002, co-localized with the multi-trait mQTL.

Mapping of QTL to the poplar genome sequence

To align the genetically mapped AFLP markers with the genome sequence, AFLP analysis was repeated for primer combinations that revealed AFLP markers located in the mQTL regions, as described in Cervera *et al.* (2001). The AFLP markers were cut out from the gel, re-amplified and sequenced. To map the AFLP and single-sequence repeat (SSR) markers (for SSR sequence information, see http://www.ornl.gov/sci/ipgc/ssr_resource.htm) onto the genome sequence (assembly version 1.0, June 2004), two slightly different approaches were used, depending on the nature of the generated sequences. The AFLP markers represented by a single sequence

were mapped with BLAST N. The BLAST output was then filtered on percentage identity over the whole length of the submitted query sequence for the AFLP sequences that were smaller than 50 bp; for those sequences that were sufficiently long (>50 bp), specificity was expected to be high enough to identify single loci. The SSR primer sequences are relatively short, but are present in pairs. This feature was used by coupling both sequences from each pair with a run of 50 N in between. These sequences were then mapped onto the poplar genome sequence with BLAST N, and the output filtered to retrieve hits that had a nearly exact identity for both SSR primers and that were located in each other's vicinity.

Identification of candidate genes

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to recover the EC numbers of all known flavonoid biosynthesis enzymes in Arabidopsis. All *viridiplantae* protein sequences for each of these EC numbers were subsequently downloaded from the European Bioinformatics Institute (EBI) website (http://srs.ebi.ac.uk/) as well as all *viridiplantae* proteins that were annotated as an *O*-, *C*-, *N*- or *S*-methyltransferase. In addition to the methyltransferase proteins, we also downloaded three Pfam HMM profiles (Methyltransf_2, PF00891; Methyltransf_3, PF01596; Omt_N, PF02409) from the Pfam database (http://www.sanger.ac.uk/Software/Pfam/). The three *p*-coumarate 3-hydroxylase (*C3H*) genes of Arabidopsis (At2g40890, At1g74540, At1g74550) were downloaded from the TAIR website (http://www.arabidopsis.org/home.html), and two *p*-coumarate 3-hydroxylase proteins (AAL47545.1 and AAL47685.1), four flavone synthase proteins (NP 199072, BAB59004, AAD39549, AAF04115), and finally also seven representative members (Q9ZTK5, Q70PR7, Q8GZU0, Q94FT4, Q8LL69, Q5Y9C6 and Q5Y9C7) of the BAHD superfamily (EC 2.3.1.–) were extracted from the EBI website.

Multiple alignments of all downloaded proteins with a given EC number or from a specific protein class were constructed with CLUSTAL W (Thompson *et al.*, 1994). From these multiple alignments, all incomplete and partial proteins were discarded and the cleaned alignments were used in the next step to build HMM profiles. All predicted proteins on the poplar assembly (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html) were screened with these constructed HMM profiles to identify the poplar homologues. The building of and screening with HMM profiles was performed using HMMER 2.3.2 (Eddy, 1998). The discrimination between 'true' homologues and 'likes' was based on the E-value scores. In all cases, a clear 'drop' in the E-value score could be detected between 'true' homologues and 'likes', indicating that gene models below this threshold did not fit the profile sufficiently well.

Acknowledgements

Supplementary Material

The following supplementary material is available for this article online:

Figure S1. MS/MS fragmentation of flavonoids in negative and positive ionization mode.

Table S1 Pearson product–moment correlation coefficients

Table S2 Contribution of mIM-obtained mQTL to flavonoid concentration variance

Table S3 mQTL results of single-trait QTL analysis

Table S4 Poplar homologues of candidate genes

Table S5 Spectral data of flavonoids **1**-**15**

Table S6 NMR data for pinobanksin 3-acetate **4**

Table S7 Number of markers used for QTL analysis

This material is available as part of the online article from http://www.blackwell-synergy.com

References

Birks, J.S. and Kanowski, P.J. (1993) Analysis of resin compositional data. *Silvae Genet.* **42**, 340–350.

Boerjan, W. (2005) Biotechnology and the domestication of forest trees. *Curr. Opin. Biotechnol.* **16**, 159–166.

Bost, G., Dillmann, C. and de Vienne, D. (1999) Fluxes and metabolic pools as model traits for quantitative genetics. I. The L-shaped distribution of gene effects. *Genetics*, **153**, 2001–2012.

Burbulis, I.E. and Winkel-Shirley, B. (1999) Interactions among enzymes of the *Arabidopsis* flavonoid biosynthetic pathway. *Proc. Natl Acad. Sci. USA*, **96**, 12929–12934.

Camacho, D., de la Fuente, A. and Mendes, P. (2005) The origin of correlations in metabolomics data. *Metabolomics*, **1**, 53–63.

Cervera, M.-T., Storme, V., Ivens, B., Gusmão, J., Liu, B.H., Hostyn, V., Van Slycken, J., Van Montagu, M. and Boerjan, W. (2001) Dense genetic linkage maps of three Populus species (*Populus deltoides*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. *Genetics*, **158**, 787–809.

Cervera, M.T., Sewell, M.M., Faivre-Rampant, P., Storme, V. and Boerjan, W. (2004) Genome mapping in *Populus*. In *Molecular Genetics and Breeding of Forest Trees* (Kumar, S. and Fladung, M., eds). New York: Food Products Press, pp. 387–410.

Cervera, M.T., Storme, V., Soto, A., Ivens, B., Van Montagu, M., Rajora, O.P. and Boerjan, W. (2005) Intraspecific and interspecific genetic and phylogenetic relationships in the genus *Populus* based on AFLP markers. *Theor. Appl. Genet.* **111**, 1440–1456.

Christensen, A.B., Gregersen, P.L., Olsen, C.E. and Collinge, D.B. (1998) A flavonoid 7-*O*-methyltransferase is expressed in barley leaves in response to pathogen attack. *Plant Mol. Biol.* **36**, 219–227.

Coppieters, W., Kvasz, A., Farnir, F., Arranz, J.-J., Grisart, B., Mackinnon, M. and Georges, M. (1998) A rank-based nonparametric method for mapping quantitative trait loci in outbred half-

sib pedigrees: application to milk production in a granddaughter design. *Genetics*, **149**, 1547–1555.

Cuyckens, F. and Claeys, M. (2004) Mass spectrometry in the structural analysis of flavonoids. *J. Mass Spectrom.* **39**, 1–15.

Deng, B.-L., Lepoivre, J.A., Lemière, G., Dommisse, R., Claeys, M., Boers, F. and De Groot, A. (1997) A practical synthesis of 3-methoxyflavones. *Liebigs Ann. Chem.* **1997**, 2169–2175.

Dixon, R.A., Achnine, L., Kota, P., Liu, C.-J., Srinivasa Reddy, M.S. and Wang, L. (2002) The phenylpropanoid pathway and plant defence – a genomics perspective. *Mol. Plant Pathol.* **3**, 371–390.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Fiehn, O. (2003) Metabolic networks of *Cucurbita maxima* phloem. *Phytochemistry*, **62**, 875–886.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnol.* **18**, 1157–1161.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* **22**, 245–252.

Grattaplaglia, D. and Sederoff, R. (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics*, **137**, 1121–1137.

Greenaway, W., May, J., Scaysbrook, T. and Whatley, F.R. (1992) Compositions of bud and leaf exudates of some *Populus* species compared. *Z. Naturforsch.* **47c**, 329–334.

Heinrich, R. and Rapoport, T.A. (1974) A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur. J. Biochem.* **42**, 89–95.

Ibdah, M., Zhang, X.-H., Schmidt, J. and Vogt, T. (2003) A novel Mg^{2+}-dependent *O*-methyltransferase in the phenylpropanoid metabolism of *Mesembryanthemum crystallinum*. *J. Biol. Chem.* **278**, 43961–43972.

Jiang, C. and Zeng, Z.-B. (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, **140**, 1111–1127.

Kacser, H. and Burns, J.A. (1973) The control of flux. *Symp. Soc. Exp. Bot.* **27**, 65–104.

Kacser, H. and Burns, J.A. (1981) The molecular basis of dominance. *Genetics*, **97**, 639–666.

Knott, S.A. and Haley, C.S. (2000) Multitrait least squares for quantitative trait loci detection. *Genetics*, **156**, 899–911.

Korol, A.B., Ronin, Y.I., Itskovich, A.M., Peng, J. and Nevo, E. (2001) Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexes of quantitative traits. *Genetics*, **157**, 1789–1803.

Li, L., Popko, J.L., Zhang, X.-H., Osakabe, K., Tsai, C.-J., Joshi, C.P. and Chiang, V.L. (1997) A novel multifunctional *O*-methyltransferase implicated in a dual methylation pathway associated with lignin biosynthesis in loblolly pine. *Proc. Natl Acad. Sci. USA*, **94**, 5461–5466.

Lohaus, G., Winter, H., Riens, B. and Heldt, H.W. (1995) Further studies of the phloem loading process in leaves of barley and spinach. The comparison of metabolite concentrations in the apoplastic compartment with those in the cytosolic compartment and in the sieve tubes. *Bot. Acta*, **108**, 270–275.

Lukačin, R., Wellmann, F., Britsch, L., Martens, S. and Matern, U. (2003) Flavonol synthase from *Citrus unshiu* is a bifunctional dioxygenase. *Phytochemistry*, **62**, 287–292.

Ma, X., Koepke, J., Panjikar, S., Fritzsch, G. and Stöckigt, J. (2005) Crystal structure of vinorine synthase, the first representative of the BAHD superfamily. *J. Biol. Chem.* **280**, 13576–13583.

Markham, K.R. and Mabry, T.J. (1975) Ultraviolet–visible and proton magnetic resonance spectroscopy of flavonoids. In *The Flavonoids* (Harborne, J.B., Mabry, T.J. and Mabry, H., eds). London: Chapman & Hall, pp. 45–77.

Martínez, O. and Curnow, R.N. (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet.* **85**, 480–488.

Morreel, K., Ralph, J., Lu, F., Goeminne, G., Busson, R., Herdewijn, P., Goeman, J.L., Van der Eycken, J., Boerjan, W. and Messens, E. (2004) Phenolic profiling of caffeic acid *O*-methyltransferase-deficient poplar reveals novel benzodioxane oligolignols. *Plant Physiol.* **136**, 4023–4036.

Orians, C.M. (2000) The effects of hybridization in plants on secondary chemistry: implications for the ecology and evolution of plant–herbivore interactions. *Am. J. Bot.* **87**, 1749–1756.

Ounaroon, A., Decker, G., Schmidt, J., Lottspeich, F. and Kutchan, T.M. (2003) (*R,S*)-reticuline 7-*O*-methyltransferase and (*R,S*)-norcoclaurine 6-*O*-methyltransferase of *Papaver somniferum* – cDNA cloning and characterization of methyl transfer enzymes of alkaloid biosynthesis in opium poppy. *Plant J.* **36**, 808–819.

Riens, B., Lohaus, G., Heineke, D. and Heldt, H.W. (1991) Amino acid and sucrose content determined in the cytosolic, chloroplastic and vacuolar compartments and in the phloem sap of spinach leaves. *Plant Physiol.* **97**, 227–233.

Rohde, A., Morreel, K., Ralph, J. *et al.* (2004) Molecular phenotyping of the *pal1* and *pal2* mutants of *Arabidopsis thaliana* reveals far-reaching consequences on phenylpropanoid, amino acid and carbohydrate metabolism. *Plant Cell*, **16**, 2749–2771.

Saslowsky, D. and Winkel-Shirley, B. (2001) Localization of flavonoid enzymes in Arabidopsis roots. *Plant J.* **27**, 37–48.

Shepherd, M., Chaparro, J.X. and Teasdale, R. (1999) Genetic mapping of monoterpene composition in an interspecific eucalypt hybrid. *Theor. Appl. Genet.* **99**, 1207–1215.

Steuer, R., Kurths, J., Fiehn, O. and Weckwerth, W. (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, **19**, 1019–1026.

Sumner, L.W., Mendes, P. and Dixon, R.A. (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, **62**, 817–836.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.

Weckwerth, W. and Fiehn, O. (2002) Can we discover novel pathways using metabolomic analysis? *Curr. Opin. Biotechnol.* **13**, 156–160.

Weckwerth, W., Loureiro, M.E., Wenzel, K. and Fiehn, O. (2004) Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl Acad. Sci. USA*, **101**, 7809–7814.

Winkel, B.S.J. (2004) Metabolic channeling in plants. *Annu. Rev. Plant Biol.* **55**, 85–107.

Winter, H., Robinson, D.G. and Heldt, H.W. (1993) Subcellular volumes and metabolite concentrations in barley leaves. *Planta*, **191**, 180–190.

Wu, S., Watanabe, N., Mita, S., Ueda, Y., Shibuya, M. and Ebizuka, Y. (2003) Two *O*-methyltransferases isolated from flower petals of *Rosa chinensis* var. *spontanea* involved in scent biosynthesis. *J. Biosci. Bioeng.* **96**, 119–128.

Yin, T.M., DiFazio, S.P., Gunter, L.E., Riemenschneider, D. and Tuskan, G.A. (2004) Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. *Theor. Appl. Genet.* **109**, 451–463.